Switching streams across ears to evaluate informational masking of speech-on-speech

Axelle Calcus[1,2], Tim Schoof[1], Stuart Rosen[1], Barbara Shinn-Cunningham[3], Pamela Souza[4]

[1] *UCL Speech, Hearing and Phonetic Sciences, 2 Wakefield Street, London WC1N 1PF, United Kingdom*

[2] *Laboratoire des Systèmes Perceptifs, Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL University, CNRS, 75005 Paris, France*

[3] *Department of Biomedical Engineering, Boston University, Boston, Massachusetts, 02215, USA*

[4] *Department of Communication Sciences and Disorders, Knowles Hearing Center, Northwestern University, 2240 Campus Drive, Evanston, Illinois 60208, USA*

Financial disclosures/Conflict of Interest:

The authors declare no conflict of interest.

The authors A. Calcus and T. Schoof would like to share first authorship.

Corresponding author: Axelle Calcus, Ecole Normale Supérieure, 29 rue d'Ulm, 75005 Paris, France.

E-mail: axelle.calcus@ens.fr

Abstract

Objectives: This study aimed to evaluate the informational component of speech-on-speech masking. Speech perception in the presence of a competing talker involves not only informational masking, but a number of masking processes involving interaction of masker and target energy in the auditory periphery. Such peripherally generated masking can be eliminated by presenting the target and masker in opposite ears (dichotically). However, this also reduces informational masking by providing listeners with lateralization cues that support spatial release from masking. In tonal sequences, informational masking can be isolated by rapidly switching the lateralization of dichotic target and masker streams across the ears, presumably producing ambiguous spatial percepts that interfere with spatial release from masking. However, it is not clear if this technique works with speech materials.

Design: Speech reception thresholds (SRTs) were measured in 17 young normal-hearing adults for sentences produced by a female talker in the presence of a competing male talker under three different conditions: diotic (target and masker in both ears), dichotic, and dichotic but switching the target and masker streams across the ears. Because switching rate and signal coherence were expected to influence the amount of IM observed, these two factors varied across conditions. When switches occurred, they were either at word boundaries or periodically (every 116 ms) and either with or without a brief gap (84 ms) at every switch point. In addition, SRTs were measured in a quiet condition to rule out audibility as a limiting factor.

Results: SRTs were poorer for the four switching dichotic conditions than for the non-switching dichotic condition, but better than for the diotic condition. Periodic switches without gaps resulted in the worst SRTs compared to the other switch conditions, thus maximizing informational masking.

Conclusions: These findings suggest that periodically switching the target and masker streams across the ears (without gaps) was the most efficient in disrupting spatial release from masking. Thus, this approach can be used in experiments that seek a relatively pure measure of informational masking, and could be readily extended to translational research.

Word count: 335 (maximum 500)

Introduction

Studies investigating listening difficulties in noise suggest that the spectral characteristics of the masker strongly influence perception of a simultaneously presented target signal. Current research (mostly concerning speech targets) broadly distinguishes two types of interference between targets and maskers based on the level of the auditory pathway at which the interference is assumed to occur: peripheral and central interference. Roughly speaking, peripheral interference is thought to stem from direct interactions of energy in the target and masker (i.e., energetic masking, EM; see Moore, 2012) or disruption of the information-carrying amplitude fluctuations in the target by the amplitude fluctuations in the masker (i.e., modulation masking, MM; Stone et al., 2012, also instantiated in the intelligibility model of Jorgensen, Ewert, & Dau, 2013, for example). Central interference accounts for all masking that cannot be attributed to spectro-temporal overlap between simultaneous auditory sources, also known as informational masking (IM; Pollack, 1975). Factors influencing susceptibility to IM relate to target/masker similarity, stimulus uncertainty, etc. (for a review, see Shinn-Cunningham, 2008). Looking back on four decades of research, studies investigating IM have utilized two very different kinds of sounds: either highly controlled, simultaneous tonal sequences, or more ecological situations where speech was presented together with competing speech[1].

Early experiments sought to isolate the contribution of IM on listeners' perception of an auditory target that was spectrally remote from, hence minimizing peripheral interference with, an interfering masker. To do so, seminal IM experiments focused on situations where a fixed-frequency, regularly repeating target tone was embedded amidst a multitone background sequence whose components fell outside of a spectral "protected region" surrounding the target (Neff, Dethlefs, & Jesteadt, 1993; Neff & Green, 1987). The first parametric study evaluating detection of a target using this design revealed rather staggering results: detection thresholds were elevated by 20 to 60 dB

[1] Although we acknowledge that the definition of IM is vague and contentious, possibly due to the different kinds of sounds used to investigate it, in this work we refer to IM as *masking that cannot be explained solely by spectro-temporal overlap at the peripheral level* (i.e., EM/MM).

compared to in quiet (Kidd et al., 1994). This confirmed the presence of masking that could not be attributed to any mechanism involving spectral overlap between target and maskers.

Later studies investigated the contribution of IM in more ecologically valid situations, where a speech signal is masked by an interfering speech stream. Because speech is a broadband signal, the presence of simultaneous speakers in a complex acoustic environment will inevitably lead to spectral overlap between the target speech and maskers. However, the energy in speech is distributed relatively sparsely over time and frequency; hence, it is generally believed that speech-on-speech EM/MM has little effect on intelligibility. Instead, most of the interference from speech-on-speech masking is thought to be due to IM (Brungart, 2001).

In a pioneering investigation of the influence of IM on the perception of speech in the presence of a competing speaker, Brungart (2001) assumed that the total masking could be split into two components, IM and EM. Yet, only the total masking could be directly measured. To do so, listeners' perception of a set of keywords constituting a meaningful command was evaluated when presented together with a competing speech masker of the same form as the target (using the CRM corpus; Bolia et al, 2000). In order to evaluate the deleterious effect of IM on performance, Brungart estimated the specific contribution of EM by comparing results to those using a speech-shaped noise (SSN) with the same long-term average spectrum as the speech masker, then subtracting it from the total amount of masking. The specific contribution of IM to speech-on-speech masking was estimated as the difference in performance when the masker was steady SSN versus when it was one or more interfering speakers with the same long-term average spectrum. Be it with only one (Brungart, 2001) or several (Brungart et al., 2001; Rosen et al., 2013; Simpson & Cooke, 2005) interfering speakers, these analyses suggest that IM dominates performance in the speech-on-speech condition. Specifically, for the same SNR, intelligibility was lower when the target sentence was masked by simultaneous speech than with either speech-shaped noise or modulated speech-shaped noise, at least for SNRs from +6 to -6 dB. This was further confirmed by the analysis of error patterns, which showed that listeners who incorrectly identified the key words from the target sentence were more likely to report words from the masker sentence than other possible words in the response set.

Whereas resorting to a "subtraction strategy" initially provided valuable insights regarding the contribution of IM to ecological cocktail-party situations, this approach has important limitations. Indeed, much of the difficulty induced by SSN actually stems from a specific type of MM in which the modulations in the masker interfere with the crucial modulations in the target. Note that the kind of MM we are discussing here, as explored most thoroughly by Stone and his colleagues (Stone, Fullgrabe, & Moore, 2012; Stone & Moore, 2014), requires the interaction of target and masker energy in the periphery. There is at least one other type of MM which likely has different properties, in which target and masker energy do not interact in the periphery, and yet the modulations in the masker appear to interfere with those in the target. It is not yet clear whether this process is important or not in speech-on-speech masking, although it has been demonstrated at least once, albeit in a very artificial situation (Kwon & Turner, 2001).

SSN thus induces at least two kinds of masking arising from peripheral interactions of target and masker, EM and MM. A speech masker induces an important amount of IM in addition to EM and MM. However, SSN is not a good model of the EM/MM induced by a speech masker, so subtraction does not provide a good estimate of the IM caused by the speech masker. First, given the spectro-temporal structure of speech, the amount of MM induced by a speech masker will be different than that induced by SSN with the same long-term average spectrum because the modulations in those two sounds are very different. Second, the spectro-temporal structure of speech allows listeners to compare the outputs of different auditory filters and group together coherent spectral information (for a review, see Shamma, Elhilali, & Micheyl, 2011), while typical SSN contains independent modulation at different frequencies. Third, speech is typically periodic, and a periodic masker (even with a dynamically changing fundamental frequency contour) leads to better perception of a speech target than a comparable aperiodic masker (Steinmetzger & Rosen, 2015). In short, a subtraction strategy computing the difference in performance for a speech masker and a SSN masker will not be an accurate estimate of the IM in speech-on-speech situations.

One approach to removing spectral overlap between broadband signals such as speech (and thus reducing EM/MM) is to process sentences using a multi-band tone vocoder and randomly allocate half of the frequency bands to the target and the other half to the masker, generating an "interleaved"

speech signal (e.g., Arbogast, Mason & Kidd, 2002). Of course, while this design isolates the contribution of IM to the perception of speech in noisy backgrounds, it also degrades the natural properties of the auditory signals. Alternatively, recent research has compared monaural speech reception thresholds (SRTs) in different types of background noise to predictions based on speech intelligibility models, and confirmed the contribution of both amplitude MM and IM in addition to large effects of EM in noisy backgrounds (Schubotz, Brand, Kollmeier, & Ewert, 2016). Yet, however useful models are, this technique does not isolate the specific contribution of IM to real-life acoustic scenes, in part because of missing spatial cues.

Another solution to minimize peripheral EM/MM is to present target and masker dichotically. As the streams are presented to opposite ears, the energy of the masker cannot interact with that of the target in the periphery; any masking observed in dichotic listening must be attributed to central interference. Unfortunately, spatial separation is an important cue that helps listeners segregate simultaneous signals and successfully parse complex auditory scenes. Indeed, the masking induced when the signal and masker are co-located is reduced when spatial cues are available to distinguish target and masker positions (Freyman, Helfer, McCall, & Clifton, 1999), a phenomenon termed spatial release from masking (SRM). Thus, while dichotic presentation removes all EM/MM, it also reduces IM, thereby making it a poor control when trying to evaluate contributions of central interference to speech intelligibility.

A new approach has recently been proposed to evaluate the contribution of IM to complex auditory scenes: creating spatially ambiguous stimuli that limit SRM (thus preserving IM) while eliminating EM. Following this approach, tonal sequences of target and masker streams were presented dichotically, but their lateralization alternated over time. In sequences of pure and complex tones, this paradigm preserved significant amounts of IM even though target and maskers were instantaneously presented in separate ears (Calcus et al., 2015). In other words, target and maskers were presented in opposite ears, both switching regularly back and forth between the ears. Switches occurred during silent intervals of the sequences to avoid interrupting any component of the target or masker streams by a change of the presentation side. Slowly switching target and masker lateralization (~0.4 Hz) did not affect listeners' detection performance, causing interference similar to that of a non-switching, dichotic

condition. However, when the switching was more rapid (~1 Hz) SRM was significantly reduced, presumably because the *perceived* spatial locations of target and masker were ambiguous when spatial changes were rapid (a form of "binaural sluggishness", Grantham & Wightman, 1978). A follow-up experiment was set up to further explore the listening strategies used in the rapidly switching condition. Listeners were presented with a monotic (only one auditory channel, either in the right or left ear) and dichotic version of the rapidly switching condition. Performance was significantly better in the monotic than dichotic version of the paradigm. This suggests that, in the switching dichotic conditions, listeners tried to spatially track the target across its changes of lateralization within the sequences. This was possible at slow, but not rapid, switching rates. In fact, in the rapidly switching dichotic condition, performance was comparable to that observed in a diotic baseline, which helped to confirm that there was negligible EM in the diotic task.

The aim of the present study was to evaluate the informational component of speech-on-speech masking by using ambiguous spatial percepts induced by presenting simultaneous dichotic streams that alternated the ears of presentation. We hypothesised that switching streams across ears would result in ambiguity in the spatial lateralization of speech streams. The resulting stimuli should have little EM, but high IM, since the competing streams would be perceived at similar, ambiguous locations, leading to interference on the speech perception task. Based on our previous work, we predicted that performance in the switching condition would be significantly poorer than in a non-switching dichotic condition, where the clear perceptual spatial separation of the competing streams reduces the amount of IM induced by the interfering stream. Additionally, we hypothesised that if performance on the switching condition approached that of the diotic condition, it would suggest that there was minimal EM in the diotic condition, and confirm that spatial ambiguity reduces SRM from IM. In addition to these three conditions (dichotic, diotic, and switching), performance was measured in a quiet condition to ensure that the target was audible in the absence of masker.

We expected factors of switching rate and signal coherence to influence the amount of IM observed (for a review, see Shamma, Elhilali & Micheyl, 2011); therefore, two parameters were varied in the switching condition. First, switches were designed to appear either at word boundaries, or at a faster, periodic rate. Switches occurring during silent intervals of the target stream (i.e., at key word

boundaries) were thought to minimise interruptions in the speech sequences. However, they were relatively far apart in time, which might limit their effect in reducing SRM. Indeed, in non-speech sequences, faster switches led to increased levels of IM (Calcus et al., 2015). Increasing the switching rate in the periodically switching conditions should not only lead to greater spatial ambiguity, but will also introduce switches within words, which we further expected to broaden spatial percepts. Both of these effects should lead to greater IM than when switching at word boundaries. Second, short silent gaps were inserted in the streams after lateralization switches, which was expected to decrease continuity of the streams, increasing IM. These two manipulations yielded four switching conditions: all combinations of word boundary or periodic switches, with or without silent gaps.

Methods

Participants

Seventeen young normal-hearing participants (21 – 32 yrs, 16 female) took part in this study. All participants were monolingual American English speakers and had normal hearing as indicated by audiometric thresholds ≤ 20 dB HL at octave frequencies from .25 to 8 kHz. Study procedures were approved by the Institutional Review Boards at Northwestern University. All participants provided informed consent and were paid for their participation.

Speech recognition task

Materials

Speech reception thresholds (SRTs) were measured in response to IEEE sentences (Rothauser et al., 1969) produced by a female talker (average F0: 256 Hz) in the presence of a competing male talker (average F0: 124 Hz). The target IEEE sentences contained five key words each. The competing speech also consisted of IEEE sentences from a different subset of sentences to ensure that each sentence was only presented to the listener once, either as a target or a masker. Both talkers were American English speakers.

The masker was presented for the duration of the target sentence, with the onset of the target and masker aligned. To ensure that the masker was long enough, two IEEE masker sentences were first concatenated and subsequently cut to the duration of the target sentence. SNRs were set by keeping the level of the masker fixed and varying the level of the target. The level of the masker was calibrated to be 65 dB SPL.

Conditions

In total, seven different conditions were tested. SRTs were measured in diotic (both target and masker in both ears), dichotic (target in one ear, masker in the other ear), and quiet conditions. In addition, four different 'switching' conditions were tested. As illustrated in Figure 1, in the switching conditions, the target and masker were presented dichotically while their lateralisation switched across ears several times throughout a sentence. Switches occurred either at word boundaries (Figure 1A) or periodically every 116 ms, and with or without a brief silent gap (84 ms) inserted following the switch point (Figure 1B and 1C). The rate of switching and duration of the gap were chosen following pilot testing. An example of sentence switching at key word boundaries would be: "The birch / canoe / slid / on the / smooth / planks", where the forward slashes indicate switch points. It should be noted that since the timing of the switches was determined by the structure of the target sentence, switches in the masker sentence did not necessarily occur at word boundaries. In total, there were five switches per sentence, switches occurring on average every 440ms (average switch rate = 2.3 Hz; or 1.93 Hz with silent gaps). For periodic switching, the resulting switch rates were 8.6 Hz (no pauses) or 5 Hz (with gaps). The target and masker sentences were tapered on and off across 5 ms at the switch points to reduce spectral splatter associated with the switches. Auditory examples of the different conditions can be found in the Supplementary Material.

In the dichotic condition, the ear receiving the target was determined randomly and independently for each trial. Performance in dichotic listening conditions is typically very good, indicated by very low SRTs (c.f. Cherry, 1953). Because the SNRs in this experiment were set by varying the level of the target while keeping the level of the masker constant, there is a possibility that audibility of the target rather than the presence of the masker in the opposite ear primarily determined

the SRTs. To rule out any audibility issues, SRTs were also obtained in quiet. Note that the SRTs - or target levels - in quiet were set in exactly the same way as for the dichotic condition, except that only the target (and not the masker) was presented to the listener; i.e., target sentences were presented monaurally and the ear in which the target was presented varied randomly for every sentence.

Procedure

Participants were seated in a soundproof booth and listened over ER-2 insert earphones (Etymotic, Elk Grove Village, IL). Stimuli were presented via a digital-to-analog converter (TDT RX6, Tucker-Davis Technologies, Alachua, FL), an attenuator (TDT PA5), and a headphone buffer (TDT HB7).

Participants were asked to repeat the female target sentences, which were composed of five keywords, to the best of their ability while ignoring the competing male talker. The experimenter scored the number of correctly repeated keywords.

The SNR was varied using an adaptive procedure (Plomp and Mimpen, 1979). Pilot data were used to set the SNRs for the first sentence in a block, ranging from -45 to -20 dB SNR for the different conditions (-45 dB SNR in quiet is equivalent to a stimulus level of 20 dB SPL). The first sentence was repeated until at least 3-5 words were correctly repeated, or the SNR reached 25 dB (6 dB increments). The SNR for each subsequent trial was decreased by 2 dB if 3-5 keywords were correctly repeated, or increased by 2 dB if 0-2 keywords were correctly repeated. SRTs were calculated as the average of all reversals for which the step size was 2 dB.

Participants received a brief practice block to familiarize themselves with the task. The practice block consisted of five sentences in the diotic condition, with the SNR for the first trial set to 0 dB. In the experiment proper, each block consisted of 20 sentences. The order of conditions, target lists, and masker lists were all randomized across participants using a Latin Square design. SRTs for all conditions were measured twice. A measurement was repeated, with a different set of target sentences, when fewer than three reversals were obtained or when the standard deviation across the reversals was larger than 4 dB. In addition, a measurement was repeated when the SRTs for the two repetitions of a

condition differed by more than 3 dB, in which case only the two SRTs that were closest in value were included in the data set.

Statistical analyses

The SRTs were analysed using linear mixed effects models in R (using the *lme()* function from the *nlme* package; R Core Team, 2016; Pinheiro et al., 2016). It is important to note that SRTs were not averaged across blocks, but instead SRTs for the two measurements for each condition were both included in the models. SRTs for one participant could not reliably be collected in the periodic switch condition with silent gaps. The assumptions for linear mixed effects models were met; Q-Q plots indicated that the residuals of all the models were normally distributed. The data set contained no outliers (mean +/- 3 sd). All significant results reported below remained significant after Bonferroni correction.

Results

The SRTs for the seven different conditions (diotic, dichotic, switching at word boundaries or periodically with and without silent gaps, and quiet) are plotted in Figure 2 (means and standard deviations are provided in Table 1).

The effect of switching dichotic speech streams across the ears on SRTs

To evaluate the informational component of speech-on-speech masking, SRTs were measured when target and masker speech streams were switched across the ears. Peripherally generated masking (EM/MM) can be eliminated by presenting the two streams dichotically. However, this also reduces IM by providing listeners with lateralization cues. We hypothesised that switching the streams across the ears would largely disrupt any spatial release from masking that would otherwise result from simple dichotic presentation, leading to higher (i.e. poorer) SRTs.

To examine the effect of switching the target and masker streams across the ears in general, the SRTs were analysed using a linear mixed effects model with one fixed factor, condition (diotic, dichotic,

switching (collapsed across the four different switching conditions), and quiet), and three random intercepts, listener, target sentence list, and masker sentence lists. Two contrasts were specified to assess the effects of switching (switch vs. dichotic, and switch vs. diotic). In addition, a contrast was specified to examine whether the SRTs in the dichotic condition were driven by audibility (quiet vs. dichotic).

The results, summarized in Table 2, indicate that performance on the speech perception task was significantly poorer when the target and masker were switched across the ears than when the stimuli were presented dichotically, with a mean difference of 5.9 dB ($p < 0.001$). In addition, performance in the switching condition was better than in the diotic condition, with SRTs on average about 17.2 dB lower when the target and masker were switched across the ears ($p < 0.001$). Furthermore, it is important to note that the SRTs in the dichotic condition were not limited by audibility, with performance in quiet significantly better (by 6.3 dB on average) than in the dichotic condition ($p < 0.001$). Target levels in quiet are provided here in terms of dB SNR (mean: -48.7, sd: 3.3) for easy comparison with the test conditions in noise. In terms of absolute presentation levels, this corresponds to an average SRT of 16.3 dB SPL (range: 10.4 – 22 dB SPL).

The effects of different switching conditions on SRTs

Switching target and masker streams across the ears increases SRTs compared to a simple dichotic presentation, which suggests that the SRM resulting from dichotic presentation is disrupted. It remains unclear, however, to what extent switching rate and the presence of silent gaps (i.e. signal coherence) influence SRTs. We hypothesized that the switching rate in the periodically switching conditions would lead to poorer performance compared to conditions when switches were only introduced at word boundaries (i.e. at a slower rate). Second, we hypothesized that short silent gaps inserted in the streams after lateralization switches would decrease continuity of the streams, resulting in higher (i.e. poorer) SRTs.

The SRTs for the different switching conditions were analysed using a linear mixed effects model to examine whether changes in switching rate (either at word boundaries or periodically) and signal continuity (with or without silent gaps) affected the IM component. The model included three fixed factors: switch rate (word boundary, periodic), gap (with, without), and the interaction term. The

model also included two random intercepts: listener and target list. Model comparisons based on the Akaike Information Criterion (AIC; Akaike, 1974) indicated that adding masker sentence list as a random factor did not improve the model fit.

The results, summarized in Table 3, indicate a main effect of switch rate, with poorer performance (by about 8.8 dB) when switches occurred periodically compared to when they occurred at word boundaries ($p < 0.001$). Similarly, there was a significant (albeit more modest) main effect of gap: SRTs were on average 2.3 dB lower (i.e., better) when silent gaps were present compared to when they were not ($p < 0.001$). There was no significant interaction between the presence of a gap and switch rate ($p = 0.2$). However, contrary to our initial expectations that silent gaps would interfere with the coherence of the target stream, thereby increasing IM, these results show that introducing gaps (both after word boundary switches and after periodic switches) made performance better.

Periodic switches without silent gaps may be most effective at inducing IM

These results suggest that the periodic switch condition without silent gaps may be most effective at inducing IM in speech-on-speech masking despite dichotic presentation, given that performance is poorest in this condition. To examine the extent to which this method could isolate the informational component of speech-on-speech masking, a follow-up linear mixed effects model was evaluated. The model treated condition (periodic switch without gaps, diotic, dichotic) as a fixed factor and included two random factors (intercept), listener and target sentence list. Model comparisons indicated that adding a random intercept for sentence list did not improve the model fit. Two contrasts were specified to assess the effects of the switching condition (switch vs. dichotic, and switch vs. diotic).

The results (see Table 4) indicate that performance in the periodic switching condition (no gaps) was significantly poorer, by about 12.2 dB, than performance in the dichotic condition ($p < 0.001$). In addition, SRTs were lower (i.e., better) than in the diotic condition (11.1 dB on average, $p < 0.001$). The general pattern of results is similar to that found when switches were introduced at word boundaries in the sense that performance on the switching condition falls in between that for the diotic and dichotic conditions (see Table 5 for additional analyses). However, the magnitudes of the effect were different. When switches occurred periodically, SRTs were closer to the diotic condition (differing by 12.2 dB in

the periodic switch condition compared to 20.8 dB in the word boundary switch condition). SRTs were further from the dichotic condition in the periodic switch condition (11.1 dB) compared to the word boundary switch condition (2.5 dB). Overall, these results suggest that switching rate may be the most important variable in reducing IM: faster switch rates may produce more ambiguous spatial percepts, causing greater reduction in SRM.

Discussion

When a target speech stream and competing distractor speech stream are presented dichotically, in opposite ears, any perceptual interference of the distractor on the target must be due to IM, not EM/MM; however, because ordinary dichotic stimulation leads the two streams to be perceived as coming from distinct locations, it also eliminates most IM. The present study aimed to evaluate the effect of introducing spatial ambiguity into dichotically presented competing speech using an ear-switching paradigm. Previous research using (discrete) pure and complex tones showed that rapidly alternating the lateralization of dichotically presented target and masker streams did not cause SRM from IM, even though instantaneously, there are lateralization cues inherent in the stimuli (Calcus et al., 2015). For these simple tone sequences, switching lateralized streams at a slow rate of only 1 Hz lead to performance comparable to when the streams were presented diotically. Not only does this result suggest that dichotic switching reduces SRM by producing spatially ambiguous percepts, it suggests that in these experiments, the perceptual interference in the diotic presentation was essentially all IM, with little or no contribution of EM/MM. Here, we implemented the switching paradigm using (continuous) speech sentences. Our main result suggests that, whereas the switching paradigm preserves more IM than a traditional non-switching dichotic condition, it does not elicit performance comparable to the diotic condition, even at the most rapid switching rate tested.

At first glance, these results appear to contradict the findings of Calcus et al. (2015) because performance in the switching speech-on-speech condition is always better than in the diotic condition. The main explanation for this discrepancy most likely lies in the nature of the material used. Indeed, thanks to a protected region surrounding the target, there was very little EM/MM in the diotic condition

of Calcus et al. (2015) for both pure and complex tones. This was not the case for speech-on-speech material used here, where the diotic condition elicited both EM/MM and IM. This combination of interference might lead to poorer performance in the diotic conditions for the speech-on-speech conditions tested here compared to the nonspeech stimuli previously tested (Calcus et al., 2015), hence leading to a larger discrepancy when comparing performance in the diotic to the switching condition. Additionally, the speech signal not only carries meaning but is also a continuous signal; the discrete tones constituting the tonal sequences used previously are separated in time. Spectro-temporal coherence of an auditory target is known to improve its detection and identification in complex auditory scenes (Shamma, Elhilali & Micheyl, 2011). The continuous nature of the speech signal increases the coherence of the auditory target, which reduces effects of the interfering speech masker, even when gaps disrupt the continuity of the speech signal. Another factor limiting the impact of the masker on the listeners' performance is the gender difference between the target and maskers used here; differences in talker gender reduce IM to the point that spatial separation between the talkers has a modest impact (Brungart, 2001). The possibility that switching has a smaller impact when target and maskers are more coherent and perceptually distinct in other perceptual dimensions could be tested in tonal sequences by decreasing the inter-stimulus interval between targets, increasing the frequency separation between target and maskers, and/or introducing timbre differences that differentiate the target from maskers.

The condition that had the greatest IM combined periodic interruptions of the sentences *without* gaps. Indeed, we observed a small but significant improvement in performance when sentence interruptions (both periodic and at word boundaries) were followed by short silent gaps. This finding does not fit our prediction that gaps would reduce target stream coherence and increase IM. This effect is likely due to the fact that the gaps decreased the switch rate and caused both the target and masker streams to be less spatially ambiguous. As a result of these changes, each individual segment of the target stream and the masker stream (between the imposed gaps) might have had more distinct spatial positions, leading to stronger SRM and limiting the deleterious effect of the switches on lateralization. However, if the decrease in switching rate was the main factor contributing to the performance improvement in the presence of silent gaps, we might expect that a larger decrease (as observed in the periodically switching sequences) would lead to a greater benefit from the gaps. This was not the case,

as we did not observe a significant interaction between switching rate and gaps. Alternatively, this improvement in performance when there were silent gaps between words could also arise because the gaps allowed listeners longer processing time, ultimately leading to modest but significant benefits in speech intelligibility (Best et al., 2015; Gygi & Shafiro, 2014). Further studies are warranted to disentangle the respective contributions of switching rate and processing time on the effect of the presence of gaps in different switching conditions. Regardless, the gaps enhanced SRM; therefore, in order to preserve significant amounts of masking, it is best to avoid gaps between the switches in lateralization.

Both switching periodically and at word boundaries led to a significant increase in SRTs compared to the dichotic condition, indicating that they could both be used to evaluate the contribution of IM to speech-on-speech situations. However, the effect size was much larger in the periodically switching condition. Even though switching at word boundaries might be more ecologically valid, periodic switching may be the most effective way to preserve IM in dichotic speech-on-speech conditions. Perhaps the most likely explanation for the poorer performance when switching periodically rather than at word boundaries lies in the fact that the switch rates are simply more rapid in the periodic conditions than in the word boundary conditions. For word boundary switches the average switch rate was 2.3 Hz (no gaps) or 1.93 Hz (with gaps), while for periodic switching the switch rates were 8.6 Hz (no gaps) or 5 Hz (with gaps). Faster switch rates likely lead to more ambiguous and/or broader spatial percepts, thereby preserving IM.

Another possible explanation for the poorer performance when switching periodically is the short loss of information at +/- 5ms around the time of the interruption (i.e., the duration of the ramp used to prevent spectral splatter), which might jeopardize lexical access of target keywords. To examine the effect of tapering, we ran a short follow-up experiment comparing SRTs in diotic and dichotic conditions (i.e., without switching) with and without tapering the streams every 116 ms (8.6 Hz). There was no significant effect of tapering in the diotic and the dichotic conditions (all ps $\geq$ 0.4). Therefore, the impact of the periodic switching cannot be explained solely by the loss of information due to tapering the signal on and off across 5 ms at the edges of the switch times. Further studies are needed

to determine whether the nature of the switching conditions (i.e., at word boundary or periodically) might contribute to the differences observed between the two switching rates tested here.

A third hypothesis is that listeners may have perceived a coherent auditory stream despite the changes in lateralization. Words bind automatically due to low-level, spectro-temporal cues, whereas binding across word boundaries requires higher-level predictability, based on cues such as location, pitch or even meaning (e.g., Kidd et al., 2013, 2014). Spatial percepts are formed by combining spatial cues across all sound elements that are bound together (e.g., see Best et al., 2007). When spatial switching occurs within a word, the perceived location of that word will be less precise and more diffuse than when the spatial cues in the entire word are consistent, thus reducing the efficacy of SRM.

What is the nature of the difficulty induced by the "switching"? Using tonal sequences, rapidly switching target and masker streams (~1 Hz) was thought to prevent listeners from accessing the spatial cues inherent to dichotic listening, hence avoiding spatial masking release. This was not the case when target and maskers switched slowly (~0.4 Hz). Specifically, listeners likely tried to spatially track the target across the changes in lateralization, which was easier to perform when switching was slower. This difficulty is thought to be related to a phenomenon of "switching sluggishness" of the auditory system (Calcus et al., 2015), based upon the phenomenon of binaural sluggishness (Grantham & Wightman, 1978). Here, using speech material, periodically switching sequences (~8.6 Hz) did limit access to spatial information when compared to a dichotic condition, but not to the extent of avoiding SRM like a diotic condition would, even though the switching rate was significantly faster than with the nonspeech material. Yet by increasing uncertainty regarding the target characteristics and increasing spatial similarity between target and masker, both of which contribute to IM (Durlach et al., 2003), the switching paradigm still preserves more IM than the non-switching dichotic condition. Moreover, periodically switching target and masker streams does successfully eliminate spectral overlap between simultaneous streams at the peripheral level while minimizing degradation of the auditory stimuli. Further studies are warranted to determine where the "switching sluggishness" comes from, be it from the binaural auditory system (e.g., Culling & Mansell, 2013) or from a higher cognitive level (e.g., Koch, Lawo, Fels, & Vorländer, 2011).

In everyday listening situations, understanding a speaker of interest amongst interfering speakers is challenging for most listeners, even those who have normal audiometric thresholds (Ruggles, Bharadwaj, & Shinn-Cunningham, 2011). Crucially, certain populations might be particularly affected by the contribution of IM to speech-on-speech listening situations, for example children (who have immature linguistic knowledge) or older adults (whose diminished cognitive resources and reduced lexical inhibition may make it difficult to focus on and process the target speech).

With regard to children, because frequency selectivity is fully mature in the first year of life (e.g., Eggermont et al., 1996), central interference is thought to account for most of the children's difficulty encountered in noisy backgrounds. This is consistent with later observations of larger effects of IM in children aged 4 to 16 years than in adults (Wightman & Kistler, 2005). This difficulty might stem from immature stream segregation mechanisms, which have been shown to develop over time (Sussman et al., 2007). Given the crucial role of attention on streaming of auditory objects (Woods & McDermott, 2015), and the existing evidence that speed and efficiency of attention allocation develop beyond the age of 12 years (Gomes, Duff, Barnhardt, Barrett, & Ritter, 2007), further research is needed to explore developmental effects of auditory attention and stream segregation on speech-on-speech performance in children, and their interplay in adults.

In the case of older adults, cognitive impairment might account for a significant proportion of the speech-on-speech difficulties (Füllgrabe, Moore, & Stone, 2015). Consistent with this possibility, specific difficulties in situations maximising IM compared to situations maximising EM/MM have been reported in normal-hearing older adults (Schoof & Rosen, 2014). This may be related to older adults' reduced lexical inhibition (Dey & Sommers, 2015; Robert & Mathey, 2007), where competing information would be expected to interfere to a larger extent. However, other studies failed to report an effect of age on speech-on-speech performance, despite observing that some older adults had particular difficulties in such situations (Agus, Akeroyd, Gatehouse, & Warden, 2009). Further research is required to shed light on the precise nature of the speech perception difficulties typically experienced by older adults.

To reconcile disparate findings and advance management options, isolating the contribution of IM to difficulties perceiving speech-on-speech is necessary to better specify the nature of the difficulties

encountered by many listeners in noisy backgrounds. The technique reported here, consisting of periodic switching of speech streams, successfully preserves IM despite dichotic presentation, while eliminating spectral overlap at the peripheral level. Therefore, this paradigm provides a useful measure of IM in noisy environments that can be readily extended to translational research.

Acknowledgments

The authors would like to thank Rachel Ellinger and Andrea Cunningham for their help with data collection. This work was supported by NIH R01 DC 60014 grant awarded to Pamela Souza, and an iCARE ITN (FP7-607139) European fellowship to Axelle Calcus.

Figure legends

Figure 1: Illustration of the main experiment conditions. A: Switching at word boundaries; B: Periodic switching; C: Periodic switching with silent gaps. For each condition, the upper panel represents the right ear channel, the lower panel represents the left ear channel.

Figure 2: Speech reception thresholds (in dB) in response to a female talker in the presence of a male talker are plotted in diotic, dichotic, switching, and quiet conditions. Results are plotted for four different switching conditions: switches were either introduced periodically or at word boundaries, and with or without gaps inserted at each switch point.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19 (6), 716–723.

Arbogast, T., Mason, C., & Kidd, G. (2002). The effect of spatial separation on informational and

energetic masking of speech. *J. Acoust. Soc. America*, *112*(5), 2086–2098.

Agus, T., Akeroyd, M., Gatehouse, S., & Warden, D. (2009). Informational masking in young and elderly listeners for speech masked by simultaneous speech and noise. J. Acoust. Soc. America, 126(4), 1926–1940.

Best, V., Gallun, F., Carlile, S., & Shinn-Cunningham, B. (2007). Binaural interference and auditory grouping. *J. Acoust. Soc. America*, *121*(2), 1070-1076.

Best, V., Mason, C. R., Swaminathan, J., et al. (2015). Does providing more processing time improve speech intelligibility in hearing-impaired listeners? Presented at the Proceedings of the Meetings on Acoustics.

Bolia, R.S., Nelson, W.T., Ericson, M.A., and Simpson, B.D. (2000). A speech corpus for multitalker communications research. *J. Acoust. Soc. America*, 107, 1065–1066.

Bronkhorst, A. (2000). The cocktail-party phenomenon: A review of research on speech intelligbility in multiple-talker conditions. *Acta Acustica, 86,* 117-128.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *J. Acoust. Soc. America*, *109*(3), 1101–1109.

Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *Jasa*, *110*(5), 2527–12.

Calcus, A., Agus, T., Kolinsky, R., and Colin, C. (2015). Isolating informational masking in both pure and complex tone sequences. *Ear and Hearing*, *36*(3), 330–337.

Cherry, C. (1953). Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. America,25*(5), 975-979.

Culling J. & Mansell, E. (2013). Speech intelligibibility among modulated and spatially distributed noise sources. *J. Acoust. Soc. America*, *133*(4), 2254–2261.

Day, A. & Sommers, M. (2015). Age-related differences in inhibitory control predict audiovisual speech perception. *Psychol Aging. 30*(3), 634-46.

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., & Kidd, G. (2003). Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity, *J. Acoust. Soc. America, 114*(1), 368–379.

Eggermont, J., Brown, D., Ponton, C., & Kimberley, B. (1996). Comparison of distortion product otoacoustic emission (DPOAE) and auditory brain stem response (ABR) traveling wave delay measurements suggests frequency-specific synapse maturation. Ear & Hearing, 17, 386-394.

Freyman, R., Helfer, K., McCall, D., Clifton, R. (1999). The role of perceived spatial separation in the unmasking of speech. *J. Acoust. Soc. Am. 106*(6), 3578-3588.

Grantham, D., & Wightman, F. (1978). Detectability of varying inter-aural temporal differences. *J. Acoust. Soc. Am. 63*, 511-523.

Gomes, H., Duff, M., Barnhardt, J., Barrett, S. & Ritter, W. (2007). Development of auditory selective attention: Event-related potential measures of channel selection and target detection. Psychophysiology, 44(5), 711-727.

Jorgensen, S., Ewert, S. D., & Dau, T. (2013). A multi-resolution envelope-power based model for speech intelligibility. *J Acoust Soc Am, 134*(1), 436-446. doi:10.1121/1.4807563

Gygi, B., & Shafiro, V. (2014). Spatial and temporal modifications of multitalker speech can improve speech perception in older adults. *Hearing Research*, *310*(c), 76–86.

Kidd, G., Jr, Mason, C. R., & Best, V. (2014). The role of syntax in maintaining the integrity of streams of speech. *J. Acoust. Soc. America*, *135*(2), 766–777.

Kidd, G., Jr, Mason, C. R., & Deliwala, P. S. (1994). Reducing informational masking by sound segregation. *J. Acoust. Soc. America*, *95*(6), 3475–3480.

Kidd, G., Mason, C. R., Streeter, T., Thompson, E. R., Best, V., & Wakefield, G. H. (2013). Perceiving sequential dependencies in auditory streams, *134*(2), 1215.

Koch, I., Lawo, V., Fels, J., & Vorländer, M. (2011). Switching in the Cocktail Party: Exploring Intentional Control of Auditory Selective Attention. Journal of Experimental Psychology: Human Perception and Performance, 307(4), 1140-1147.

Kwon, B. J., & Turner, C. W. (2001). Consonant identification under maskers with sinusoidal modulation: Masking release or modulation interference? *Journal of the Acoustical Society of America*, *110*(2), 1130-1140.

Moore, B. (2012). Frequency selectivity, masking and the critical band. In Moore, B.C.J., *An*

Introduction to the Psychology of Hearing (6th ed., pp. 67-101). Bingley, UK: BRILL.

Neff, D. L., & Green, D. M. (1987). Masking produced by spectral uncertainty with multicomponent maskers. *Perception & Psychophysics, 41*(5), 409–415.

Neff, D. L., Dethlefs, T. M., & Jesteadt, W. (1993). Informational masking for multicomponent maskers with spectral gaps. *J. Acoust. Soc. America, 94*(6), 3112–3126.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2016). *nlme: Linear and nonlinear mixed effects models*. R package version 3.1-127, URL: http://CRAN.R-project.org/package=nlme

Plomp, R., and Mimpen, A. (1979). Improving the reliability of testing the speech reception threshold for sentences. *Audiology* 18, 43–52.

Pollack, I. (1975). Auditory informational masking. *J. Acoust. Soc. Am. 57(suppl. 1),* S5.

R Core Team (2016). *R: A language and environment for statistical computing*. R version 3.3.0. R Foundation for Statistical Computing, Vienna, Austria. URL: www.R-project.org

Robert, C. & Mathey, S. (2007). Aging and lexical inhibition: the effect of orthographic neighborhood frequency in young and older adults. *J Gerontol B Psychol Sci Soc, 62*(6),340-342.

Rosen, S. (1992). Temporal Information in Speech: Acoustic, Auditory and Linguistic Aspects. *Philosophical Transactions: Biological Sciences, 336*(1278), 367–373.

Rosen, S., Souza, P., Ekelund, C., & Majeed, A. A. (2013). Listening to speech in a background of other talkers: Effects of talker number and noise vocoding, *J. Acoust. Soc. America, 133*(4), 2431–2443.

Rothauser, E. H., Chapman, N. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., et al. (1969). IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust*. 17, 225–246.

Ruggles, D., Bharadwaj, H. & Shinn-Cunningham (2011). Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. *Proceedings of the National Academy of Science, 103*(37), 15516-15521.

Schubotz, W., Brand, T. Kollmeier, B., & Ewert, S. (2016). Monaural speech intelligibility and detection in maskers with varying amounts of spectro-temporal speech features. *J. Acoust. Soc. Am. 140*(1), 524-540.

Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene

analysis, *Trends in Neurosciences, 34*(3), 114–123.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention, *Trends in Cognitive Sciences, 12*(5), 182–186.

Simpson, S. A., & Cooke, M. (2005). Consonant identification in N-talker babble is a nonmonotonic function of N. *J. Acoust. Soc. America*, *118*(5), 2775–2778.

Steinmetzger, K. & Rosen, S. (2015). The role of periodicity in perceiving speech in quiet and in background noise, *J. Acoust. Soc. America, 138,* 3586-3599.

Stone, M. A., Füllgrabe, C., & Moore, B. C. J. (2012). Notionally steady background noise acts primarily as a modulation masker of speech, *J. Acoust. Soc. America, 132*(1), 317–326.

Stone, M. A., & Moore, B. C. (2014). On the near non-existence of "pure" energetic masking release for speech. *J Acoust Soc Am, 135*(4), 1967-1977.

Sussman, E., Wong, R., Horváth, J., Winkler, I., & Wang, W. (2007). The development of the perceptual organization of sound by frequency separation in 5–11-year-old children. *Hearing Research, 225*(1-2), 117–127.

Wightman, F. & Kistler, D. (2005). Informational masking of speech in children: Effects of ipsilateral and contralateral distracters. *J Acoust Soc Am, 118*(5), 3164-3176.

Woods, K. & McDermott, J. (2015). Attentive tracking of sounds sources. *Current Biology, 25*(17), 1-9.